Measuring Weight Specific Quality of Life in Adolescents: An Examination of the Concurrent Validity and Test Re-Test Reliability of the WAItE

Yemi Oluboyede, PhD and Tomos Robinson, PhD

Health Economics Group, Institute of Health and Society, Newcastle University, Newcastle-upon-Tyne, UK

Keywords: adolescence, obesity, WAItE, CHU-9D, YQOL-W, health related quality of life

Text pages: 16

Tables: 3

Figures: 3

Word Count: 4000

Acknowledgements: This study was funded by the Health Foundation, project reference C0150R4035. The publication of the study results was not contingent on the sponsor's approval or censorship of the manuscript. This study was approved by the Faculty of Medical Sciences Research Ethics Committee, part of Newcastle University's Research Ethics Committee (project reference 1262/12643). Both authors report no conflict of interests. Data and Stata do-files available from the authors on request.

Abstract

Objectives: The purpose of this study was to examine the test-retest reliability of the Weight-specific Adolescent Instrument for Economic-evaluation (WAItE) and its concurrent validity, as compared to the generic, preference based Child Health Utility 9D (CHU-9D) and the weight-specific Youth Quality of Life- Weight (YQOL-W).

Methods: An online survey was used to administer the three instruments on a sample of adolescents (aged 11-18). Individual responses were converted into either utility scores (CHU-9D) or health related quality of life scores (WAItE and YQOL-W). A 10 % sub-sample of the respondents also completed the WAItE one week after completion to assess test re-test reliability.

Results: 1,000 adolescents completed the online survey. There was a strong correlation between the WAItE and both the CHU-9D (0.729, $P<0.001$) and the YQOL-W (0.750, $P<0.001$). All three instruments were able to discriminate according different weight status categories and a measure of self-assessed health. Unlike the CHU-9D or YQOL-W, the WAItE did not show a substantial ceiling effect. The WAItE also showed acceptable levels of test-retest reliability.

Conclusions: The study results are encouraging, and illustrate that the WAItE can be used to reliably and accurately measure weight specific outcomes in the younger population. The WAItE can also be used to assess outcomes in cost-effectiveness analysis of weight management interventions for young people, given the instrument is less likely to display ceiling effects and may thus may be more sensitive in measuring change that results from interventions developed for this population.

1. Background

Obesity has been described as an escalating global epidemic by the World Health Organisation [1], and is estimated to cause around 3.4 million deaths per year worldwide [2]. As with the adult population, the prevalence of obesity in adolescence (approximately between the ages of 11-18 years) has been increasing in recent years [3, 4]. This raises public health concerns, as it has been shown that paediatric obesity tracks into adulthood [5]. Obesity has been associated with negative consequences that have both immediate and long-term implications on both health and health related quality of life (HRQoL) [6].

Several obesity prevention initiatives have been established by public sector organisations, such as the National Child Measurement Programme, managed by Public Health England. Furthermore, in 2006 the National Institute for Health and Care Excellence (NICE) developed the first national guidelines on the prevention, identification, assessment and management of obesity in adults and adolescents in England and Wales [7]. These guidelines recommended several interventions for the prevention and management of obesity in young people, including interventions related to lifestyle, behaviour, physical activity and diet. However, although such public health interventions may be effective in reducing levels of adolescent obesity, it is becoming increasingly important for efficient resource allocation to also assess the value for money of such interventions [8].

Cost utility analysis (CUA) is commonly used to inform whether new interventions should be made available within a publicly funded healthcare system. In CUA, benefits are commonly measured using quality adjusted life years (QALYs), which take into account both the length of life and the HRQoL of the patients. Preference based measures (PBMs) can be used to calculate an individual utility score, where a value of 1 represents full health and a value of 0 represents death. Preference-based measures differ from non-preference-based measures in the way the scoring algorithms have been derived. Specifically, these scoring algorithms are estimated from the values that patients place upon different aspects of health, instead of being a summative scoring procedure [9]. Although generic PBMs are in theory comparable across multiple clinical areas, condition specific PBMs may be more sensitive to certain disease-specific improvements, and therefore may be a better way of valuing patient benefit.

Within the obesity literature, a number of weight-specific HRQoL instruments have been developed for use in adolescence, including the KINDL-obesity module [10], the IWQOL-Kids [11], the M-A-QoL Q [12], Sizing Me Up [13], YQOL-W [14] and OPOI [15].

However, none of these tools are preference based, nor were they designed for this purpose. Furthermore, to date only the YQOL-W has been fully validated. In response to this, a weight specific HRQoL instrument for adolescents, the Weight-specific Adolescent Instrument for Economic-evaluation (WAItE), has been developed. The WAItE is a brief 7-item measure incorporating the views and experiences of adolescent girls and boys aged 11-18 years (for full details of the development of the WAItE please see Oluboyede et al. [16]). The WAItE was developed to be feasible for a valuation study to be conducted, in order to derive preference weight to operationalise the calculation of QALYs, and is currently the only weight specific tool for adolescents with this quality. To date, a valuation study has yet to be carried out. It is important to conduct a robust validation of the WAItE, in order to provide evidence of its criterion validity and reliability to endorse future use.

With this context, the first aim of this study was to examine the concurrent validity of the WAItE, as it is imperative to evaluate the performance of the WAItE against previously validated HRQoL tools [17]. The second aim of this study was to evaluate the test-retest reliability of the WAItE, in order to ensure that the WAItE is free of measurement error and gives consistent responses over two time-points where no change in weight status has been observed [18].

2. Methods

Sample

An online survey was developed with the survey company Survey Sampling International (SSI) for administration to a sample of adolescents aged 11-18 residing in the United Kingdom (UK). The survey contained two sections. The first section comprised of a series of sociodemographic questions including age, gender, height, weight, self-assessed weight status and self-assessed health. The second section comprised the WAItE, CHU-9D and YQOL-W instruments, randomised in order.

A sample of around 15,000 adult participants with children between the ages of 11-15 were approached to complete the survey from SSI's large participant panel. The 11–15 year old participants were then able to complete the survey, given the consent of their guardian. Furthermore, around 2500 16 to 18 year olds were directly invited to complete the survey by SSI. The survey was left open until 1000 participants had completed the survey. Respondents to SSI surveys receive an average of £0.30 per 5 minute interview. The median time to complete the survey was 6 minutes.

In order to examine the test-retest reliability of the WAItE, a 10% sub-sample of the respondents were contacted again around one week later, and completed a shortened version of the questionnaire for a second time. This one week time period was based upon a similar study examining the test-retest reliability of the YQOL-W [14]. The vast majority of those contacted again completed the shortened survey within 10 days, with the maximum time between the completion of the two surveys being 18 days (n=2). The survey was completed between July 5th and September 25th, 2017, and was approved by Newcastle University's Faculty of Medical Sciences Research Ethics Committee (project reference 1262/12643).

The concurrent validity of the WAItE as compared to the CHU-9D and YQOL-W was assessed in terms of its ability to differentiate between individuals of different weight statuses. The CHU-9D was chosen as a comparator as it is a generic, preference based HRQoL tool commonly used in economic evaluation, while the YQOL-W was chosen as a comparator as it was the only paediatric weight specific HRQoL tool available at the beginning of the study. A quota on weight status was stipulated initially, with an aim of having the 1,000 respondents split equally between three weight groups: normal, overweight and obese. The three weight groups were formed based on the respondent's self-reported height and weight, and the age specific cut-off points established by Cole et al. [19]. Adolescents in the 85th percentile of the age and gender adjusted weight distribution were considered overweight, and those in the 95th percentile were seen as being obese.

HRQoL Scoring and Utility Measurement

WAItE

The WAItE has seven dimensions (relating to tiredness, walking, participation in sports, concentration, embarrassment, unhappiness and being treated differently), with a five-level frequency response scale representing the increasing degrees of severity (ranging from 'never' to 'always'). The WAItE total score is calculated by simply summing the answers of the seven dimensions, and is scored between 7 and 35 [16]. In analysis, this WAItE 'total score' was reverse coded so that a higher WAItE 'total score' indicated a higher quality of life, in line with the CHU-9D.

CHU-9D

The CHU-9D has nine dimensions (related to being worried, sad, in pain, tired, annoyed, schoolwork, sleep, daily routine, and ability to join in activities), each with a five-level

frequency response scale representing increasing degrees of severity (ranging from, for example, 'I don't feel worried today' to 'I feel very worried today') [20]. The instrument has been validated for use in adolescent populations. We used the scoring algorithm based on the preferences of the UK adult general population, using the syntax provided by the authors. These utility scores have a minimum value of 0.33 and a maximum value of 1.

YQOL-W

The YQOL-W was developed as a measure of weight-specific QOL in youth, and was based on over 50 interviews with adolescents living in the USA and Mexico [14]. The YQOL-W has 21 dimensions (including those related to depression, exercise and social anxiety) each with 11 different levels of severity (ranging from 'not at all' to 'very much'). The YQOL-W can be reverse coded and converted into a total score between 0-100, with 100 being the maximum score.

Statistical Analysis

To test the concurrent validity of the WAItE, we firstly compared its statistical properties with the generic, preference based CHU-9D and the weight specific YQOL-W. Descriptive analysis, including means, standard deviations and medians, were initially estimated. The distribution of the CHU-9D utility score and the WAItE and YQOL-W HRQoL total scores were tested for normality using the Shapiro-Francia test [21], and the non-parametric Spearman rank correlation coefficient was used to assess the level of agreement between the instruments [22]. In line with Cohen [23], a coefficient over 0.8 was seen to indicate a high level of correlation, and a coefficient of 0.6-0.8 was seen to indicate a good level of correlation. Additionally, Bland-Altman plots [24] were used to study the limits of agreement between the WAItE and both the CHU-9D and YQOL-W. We hypothesised that the highest correlation would be seen between the WAItE and the YQOL-W, as both these instruments are weight specific rather than generic.

The concurrent validity of the WAItE was further assessed by analysing the performance of the WAItE compared to the CHU-9D and YQOL-W in its ability to discriminate between gender and age adjusted BMI categories, the five levels of self-reported general health, the six levels of self-reported weight status and the presence or absence of a long-term illness or disability. It was expected that respondents with better health statuses (for example those who reported themselves to have 'excellent' or 'good' health) would have a higher HRQoL than those who had worse health statuses. Due to the non-normal distribution of all three

instruments, two nonparametric tests (the Kruskal-Wallis test [25] and the Mann-Whitney U Test [26]) were adopted to compare the respective utilities and HRQoL scores between the various subgroups.

To examine the test-re-test reliability of the WAItE, we utilised a 10% random sub-sample who completed the WAItE again one week later, and compared the responses with those from the full estimation sample. The test-re-test reliability of the WAItE was assessed using methods recommended by the Consensus-based Standards for the Selection of Health Measurement Instruments checklist manual [27]. To assess the test-re-test reliability of the continuous WAItE total score, the intraclass correlations coefficient (ICC) was used [28]. This measure is seen to be superior to measures such as the Spearman rank correlation coefficient, as it takes into account the possibility of systematic error. Results from individual mixed effects models are presented, as the 'raters' (the survey participants) were fixed over time [29]. The guidelines of Cicchetti [30] were used to assess the degree of agreement using the ICC. An ICC <0.4 was seen to indicate poor agreement, an ICC between 0.41-0.6 was seen to indicate fair agreement, an ICC between 0.61 and 0.74 was seen to indicate good agreement and an ICC > 0.75 was seen to indicate almost perfect agreement.

To assess the test-re-test reliability of the ordinal, individual scales of the WAItE, weighted Kappa coefficients were used. This measure is seen to be superior to the standard Kappa coefficient or the proportion agreement, as it takes into account the possibility of chance agreement [31]. Results with both linear and quadratic weights are presented. In line with Cannaway and Frew [32], Landis and Koch's [33] guidelines were used to assess the degree of agreement using Kappa coefficients. A Kappa coefficient <0.2 indicates poor agreement, a coefficient between 0.21-0.40 indicates fair agreement, a coefficient between 0.41-0.6 indicates moderate agreement, a coefficient between 0.61-0.8 indicates substantial agreement, and a coefficient >0.81 indicates almost perfect agreement. All statistical analysis was undertaken using Stata v14.1 [34].

3. Results

Comparison of the WAItE, CHU-9D and YQOL-W

A total of 1,000 participants completed the online survey. The survey did not allow the respondents to answer the next question in the survey until the current question had been fully completed, and therefore there was no missing data on any of the key variables. However, there were 25 respondents excluded from the final sample due to unfeasible body mass index

(BMI) values. The final sample included more respondents classed as obese (n=401, 41.13%) and normal weight (n=353, 36.21%) than overweight (n=221, 22.66%), although there were still sufficient numbers in the three groups to conduct a robust empirical analysis. As displayed in Table 1, the mean age of the respondents was 15.4 years, and 50.6% of the respondents were female. 22.5% of the adolescents reported their health as either 'fair' or 'poor', while 24.4% of the adolescents reported themselves as being 'moderately overweight' or 'very overweight'. 57.4% of the adolescents reported themselves as having some form of illness or disability.

[TABLE 1 ABOUT HERE]

The distributions of the WAItE, CHU-9D and YQOL-W are displayed in the online supplementary materials. The mean WAItE score was 25.39, while the median score was 26. The WAItE did not display a substantial ceiling effect, with only 3.1% of the responses reporting the maximum WAItE total score of 35. The mean utility of the CHU-9D was 0.81, while the median utility was 0.83. The distribution of the CHU-9D utility score was negatively skewed and showed evidence of having a ceiling effect, with 20.1% of the respondents having a reported utility of over 0.95, and 9.6% of the respondents reporting a maximum utility value of 1. Previous studies using the CHU-9D have also reported this ceiling effect [32, 35, 36]. The mean YQOL-W score was 69.31, while the median score was 79.52. The YQOL-W score was extremely negatively skewed and showed evidence of a substantial ceiling effect, with 21.2% of the respondents reporting the maximum total score of 100.

Table 1 also displays WAItE, CHU-9D and YQOL-W values according to key sociodemographic variables and health status. Across the WAItE, CHU-9D and YQOL-W, statistically significant differences were observed for age and gender, with girls and older adolescents reporting a lower level of HRQoL than boys and younger adolescents respectively. All three instruments were also able to discriminate according to varying levels of self-reported weight and health statuses, as well as those who reported themselves as living with a long-standing illness or disability. For example, those respondents who self-reported their weight as being 'about right' had an average WAItE Total Score of 28.26, compared to an average WAItE Total Score of 19.66 for those who reported their weight as being 'very

overweight'. Additionally, those respondents who self-reported their health as being 'excellent' had an average WAItE Total Score of 29.95, compared to an average WAItE score of 20.48 for those who reported their health as being 'poor'.

[FIGURE 1 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

Figures 1 and 2 display scatter plot comparisons between both the WAItE and the CHU-9D and the WAItE and the YQOL-W respectively, with clear evidence of positive correlation in both cases. The Spearman rank correlation coefficient between the WAItE total score and the CHU-9D utility score was 0.731, while the correlation between the WAItE total score and YQOL-W total score was 0.747. The limit of agreement of the instruments was explored using a Bland-Altman scatter plot [24]. On average, only 4.9% of the respondents were outside the 95% limits of agreement when comparing the WAItE with the CHU-9D, and 5.7% of the respondents lay outside the 95% limits of agreement when comparing the WAItE with the YQOL-W, implying a high level of agreement in both cases.

The correlations between the individual dimensions of the WAItE and the CHU-9D and YQOL-W dimensions are displayed in the supplementary materials. Tiredness, a common dimension between the WAItE and the CHU-9D, has the highest correlation ($r = 0.562$) when comparing those instruments. Several others dimensions also had relatively large correlations, particularly those related to the 'work' and 'activity' CHU-9D dimensions. However, several dimensions between the two instruments were only modestly correlated. For example, the 'sports' WAItE dimension was relatively poorly correlated with several CHU-9D dimensions, including 'pain' ($r = 0.231$) and sleep ($r = 0.228$), and the 'tired' WAItE dimension was relatively poorly correlated with the 'sleep' ($r = 0.226$) and work ($r = 0.227$) CHU-9D dimensions.

The strength of correlation between the individual dimensions of the WAItE and YQOL-W was in general larger than the strength of correlation between the individual dimensions of the WAItE and the CHU-9D. The highest correlations were between the 'embarrassment' and 'unhappy' WAItE dimensions and various measures of numerous YQOL-W dimensions, including a dimension relating to difficulties finding appropriate clothes ($r = 0.632$ and 0.612)

and a dimension related to being ashamed (r = 0.634 and 0.595). However, not all dimensions were as strongly correlated as these examples. This was particularly evident for the 'tired' dimension of the WAItE, which had particularly low correlations with the YQOL-W dimensions related to inclusion (r = 0.165) and difficulties finding employment (r = 0.161).

Test-re-test Reliability

Tables 2 and 3 display the results for tests for the test-re-test reliability of the WAItE.

[TABLE 2 ABOUT HERE]

[TABLE 3 ABOUT HERE]

[FIGURE 3 ABOUT HERE]

As displayed in Table 2 and intuitively in the scatter plot of Figure 3, the WAItE total score showed a high level of test-re-test reliability, with individual ICCs above the 0.75 threshold value outlined by Ciccheti [30] to indicate excellent reliability. As shown in Table 3, in general the WAItE dimension scores also showed acceptable levels of test-re-test reliability, with percentage agreements ranging between 86.86% and 90.21%, and all weighted Kappa coefficients either indicating moderate or substantial agreement, depending on whether the linear or quadratic weighting strategy was used. We also conducted a 'sense check' of the reliability of the self-reported weight status measure. Although there were some instances where adolescents reported movements between weight statuses between the two time points, the measure was generally stable (all weighted kappa statistics indicated substantial agreement irrespective of which weighting strategy was used).

4. Discussion and Conclusions

This article has presented the methods and findings from an empirical study analysing the concurrent validity and test-re-test reliability of the WAItE. This is the first study to compare the WAItE with other validated HRQoL instruments, and the first study to formally assess the reliability of the WAItE. Ideally, the performance of the WAItE would have been tested against a well-established weight specific preference based tool which followed the gold standard of instrument development [37]. The CHU-9D and YQOL-W were seen as being the best comparators to test the performance of the WAItE against, given that another instrument which follows these gold standards does not yet exist.

First, the findings of this study demonstrate that the WAItE has good levels of concurrent validity, indicated by the high correlation of the WAItE with both the CHU-9D and the YQOL-W, and also the comparability of WAItE to both instruments in its ability to accurately differentiate between both general and weight specific health statuses. This indicates that the same latent construct of HRQoL is being measured in the WAItE and the previously validated measures. As expected, the correlation between the WAItE and the YQOL-W was higher than the correlation between the WAItE and CHU-9D, almost certainly due to the fact that the WAItE and YQOL-W were designed to specifically measure weight specific HRQoL, while the CHU-9D is a generic measure. Second, the findings of this study demonstrate that the WAItE overall displays acceptable levels of test-re-test reliability (ICC=0.795) indicating that the WAItE is stable and reliable over time. As a point of comparison, the YQOL-W has previously been shown to have an ICC of 0.77 [14].

As well as showing good levels of concurrent validity in relation to the CHU-9D and YQOL-W, it appears that the WAItE may have a significant advantage compared to both of these measures, as it does not exhibit a substantial ceiling effect. A substantial ceiling effect can indicate limited content validity and a reduced reliability [38], and can be considered a problem if 15-20% of respondents achieve the best possible score [39]. Given the lack of a ceiling effect, the WAItE may be more appropriate when measuring weight specific adolescent HRQoL, as it is more likely to pick up meaningful changes in HRQoL, particularly at the higher end of the distribution.

This study has several limitations. Firstly, the original aim was to have the sample equally split between normal weight, overweight and obese adolescents, in order to conduct a robust comparison between the three groups. Ultimately, the final estimation sample included 353 (36.2%) adolescents with a ‘normal’ weight, 221 overweight adolescents (22.7%) and 401 obese adolescents (41.1%), and therefore cannot be seen to be truly representative of the UK population. Recent research indicates that around 35% of adolescents can be considered either overweight or obese in the UK [40], compared to 63.8% in our sample respectively. Self-reported health was also lower than one would expect, with 22.3% of the adolescent reporting their health as either ‘fair’ or ‘poor’. Relatively recent nationally representative figures from the UK Household Longitudinal Study have shown that this figure is more likely to be around 7% [41]. Furthermore, an unusually high number of respondents in our sample (57.2%) reported themselves as having some form illness or disability, compared to a national average of around 12%, as reported in the 2015 Labour Force Survey [42]. Although the

overrepresentation overweight and obese adolescents in the sample meant that a higher than average number of individuals reporting some form of illness or disability was expected, this was still a surprising finding. However, the purpose of the study was not to be representative of objectively measured weight status according to the UK population. The oversampling of overweight and obese adolescents in the sample was seen to be necessary, seeing as a principal aim of the study was to identify if the HRQoL instruments were able to differentiate between individuals with different weight statuses, and therefore a certain number of individuals were needed in each sub-group to conduct a robust statistical analysis. Additionally, our sample can be seen as more representative of the individuals who may benefit from a weight management intervention, and therefore can be considered the population of interest.

The limitations associated with the accuracy of self-reported measures of height and weight should be also noted [43], however it is usual for measurements of weight status to rely on self-report in the current field of study. Furthermore, as with any study using online data collection methods, there is no guarantee that the respondent completed the survey themselves, and there is also the potential for both sample-selection bias and non-response bias resulting from the survey sampling methods. Finally, there is also a risk that the one week retest period may have been too short, therefore biasing the results through memory effects.

Planned future research includes the administration of the WAItE to different population groups and the estimation of a mapping algorithm to map between the weight specific WAItE and the generic, preference based CHU-9D. This mapping algorithm will enable researchers to generate a valid prediction of the CHU-9D utility score from the WAItE, and therefore facilitate an indirect estimate of QALYs using the WAItE. However, as argued by Brazier et al., [44] 'a mapping exercise is always a second best exercise compared to either the direct use of [a generic PBM] or a valuation of the condition-specific instrument'. Therefore, further planned research also includes the generation of preference weights using a representative UK population, in order to calculate QALYs, as per NICE recommendations [6]. Currently, the WAItE is a reliable and valid tool that would be appropriate for direct use in the measurement of health outcomes in public health or weight management interventions. The generation of a WAItE utility score would also enable assessment of cost-utility analysis of weight-management interventions and facilitate further comparisons between the WAItE and the CHU-9D.

Overall, the findings of this study are encouraging, and illustrate the potential for the WAItE to be used when measuring weight-specific HRQoL in adolescence. The WAItE therefore may be a key tool for decision makers now and in the future in the evaluation of weight management services and for organisations who run weight management services.

References

1. World Health Organisation (WHO). Controlling the global obesity epidemic . Available at http://www.who.int/nutrition/topics/obesity/en/ [Accessed April 20th, 2017]

2. Ng, M., Fleming, T., Robinson, M., Thomson, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet. 2014;384:766-81.

3. Flegal, K.M., Carroll, M.D., Kit, B.K., Ogden, C.L. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. J Am Med Assoc. 2012;307:491-97.

4. Ogden, C.L., Carroll, M.D., Kit, B.K., Flegal, K.M. Prevalence of childhood and adult obesity in the United States, 2011-2012. J Am Med Assoc. 2014;311:806-14.

5. Singh, A.S., Mulder, C., Twisk, J.W., Van Mechelen, W., Chinapaw, M.J. Tracking of childhood overweight into adulthood: a systematic review of the literature. Obes Rev. 2008;9(5):474-488.

6. Søltoft F, Hammer M, Kragh N. The association of body mass index and health-related quality of life in the general population: data from the 2003 Health Survey of England. Qual Life Res. 2009;18:1293-1299

7. National Institute for Health and Clinical Excellence. Obesity: the prevention, identification, assessment and management of overweight and obesity in adults and children. 2006.

8. Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the economic evaluation of health care programmes (4th ed). Oxford University Press, 2015.

9. Brazier, J., Ratcliffe, J., Saloman, J., Tsuchiya, A. Measuring and valuing health benefits for economic evaluation (2nd ed). Oxford University Press, 2017

10. Ravens-Sieberer U, Redegeld M, Bullinger M. Quality of life after in-patient rehabilitation in children with obesity. International Journal of Obesity & Related Metabolic Disorders. 2001;25:S63

11. Kolotkin, R.L., Zeller, M., Modi, A.C., et al. Assessing weight-related quality of life in adolescents. Obesity. 2006; 14:448-57.

12. Moorehead, M.K., Ardelt-Gattinger, E., Lechner, H., Oria, H.E. The validation of the Moorehead-Ardelt quality of life questionnaire II. Obesity Surgery. 2003;13:684-92.

13. Zeller MH, Modi AC. Development and Initial Validation of an Obesity-specific Quality-of-life Measure for Children: Sizing Me Up. Obesity. 2009;17:1171-77.

14. Morales, L.S., Edwards, T.C., Flores, Y., Barr, L., Patrick, D.L. Measurement properties of a multicultural weight-specific quality-of-life instrument for children and adolescents. Qual Life Res. 2011;20:215-24.

15. Doyle S. Development of a new health related quality of life instrument for use in paediatric obesity. Obes Rev. 2011;12:77-78.

16. Oluboyede Y, Hulme C, Hill A. Development and refinement of the WAItE: a new obesity-specific quality of life measure for adolescents. Qual Life Res. 2017:26: 2025-2039.

17. Murphy KR, Davidshofer CO. *Psychological testing: Principles & applications* (4th ed). Prentice-Hall, 1998

18. van Agt HME, Essink-Bot M-L, Krabbe PFM, Bonsel, G.J. Test-retest reliability of health state valuations collected with the EuroQol questionnaire. Soc Sci Med. 1994;39:1537-44.

19. Cole TJ, Freeman JV, Preece MA. Body mass index reference curves for the UK, 1990. Arch Dis Child. 1995;73:25-29.

20. Stevens K. Developing a descriptive system for a new preference-based measure of health-related quality of life for children. Qual Life Res. 2009;18:1105-13.

21. Shapiro SS, Francia RS. An approximate analysis of variance test for normality. J Am Stat Assoc. 1972;67:215-16.

22. Conover, WJ. Practical Nonparametric Statistics (3rd ed). Wiley, 1999

23. Cohen J. Statistical power analysis for the behavioural sciences (2nd ed). Routledge, 1988

24. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet. 1986;327:307-10.

25. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc . 1952;47:583-621.

26. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947:50-60.

27. Mokkink LB, Terwee CB, Knol DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol. 2006;6:2.

28. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30-46

29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420.

30. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assessment. 1994;6:284.

31. Cohen J. Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213.

32. Canaway AG, Frew EJ. Measuring preference-based quality of life in children aged 6–7 years: a comparison of the performance of the CHU-9D and EQ-5D-Y—the WAVES Pilot Study. Qual Life Res. 2013;22:173-83

33. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977:159-74.

34. StataCorp LP. Stata Statistical Software: Release 14: College Station, TX. 2015

35. Chen G, Flynn T, Stevens K, et al. Assessing the health-related quality of life of Australian adolescents: an empirical comparison of the child health utility 9D and EQ-5D-Y instruments. Value Health. 2015;18:432-38.

36. Stevens K. Valuation of the child health utility 9D index. Pharmacoeconomics. 2012;30:729-47.

37. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res. 2013;22:1889-905.

38. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34-42.

39. McHorney, CA, Tarlov, AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? Qual Life Res. 1995;*4*:293–307.

40. Niblett P and the Lifestyle Statistics Team. Statistics on Obesity, Physical Activity and Diet. Health and Social Care Information Centre London, 2015.

41. Siegler RBV. Measuring national well-being-exploring the well-being of children in the UK. 2014.

42. Office for National Statistics (ONS). Labour Force Survey. Available at: https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/labourforcesurveylfs. [Accessed November 3, 2017].

43. Rowland ML. Self-reported weight and height. Am J Clin Nutr. 1990;52:1125-33.

44. Brazier, J.E., Kolotkin, R.L., Crosby, R.D. and Williams, G.R. Estimating a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument from the SF-6D. Value Health. 2004;7:490-98.

Table 1- Participant Characteristics				
Characteristic	***n* (%)**	**WAItE Total Score**	**CHU-9D Utility Score**	**YQOL-W Quality of Life Score**
Full Sample (%)	975 (100%)			
Mean (SD)		25.39 (5.99)	0.81 (0.14)	69.31 (30.85)
Median (IQR)		26 (21-30)	0.83 (0.70-0.92)	79.52 (44.29-98.57)
Gender				
Males	482 (49.4%)	26.25 (5.72)	0.82 (0.14)	73.25 (31.10)
Females	493 (50.6%)	24.55 (6.14)	0.79 (0.15)	64.89 (30.07)
P - Value	-	<0.001	<0.001	<0.001
Age (Group)				
11-15	361 (37.0%)	28.60 (5.43)	0.88 (0.12)	82.19 (25.79)
16-18	614 (63.0%)	23.50 (5.48)	0.76 (0.14)	61.29 (30.97)
P - Value	-	<0.001	<0.001	<0.001
Weight Status				
Normal	353 (36.2%)	27.10 (5.64)	0.84 (0.14)	82.18 (24.17)
Overweight	221 (22.7%)	26.10 (5.67)	0.81 (0.14)	73.23 (26.85)
Obese	401 (41.1%)	23.49 (5.93)	0.78 (0.14)	55.12 (32.41)
P - Value	-	<0.001	<0.001	<0.001
Self-Assessed Weight				
Very Overweight	97 (10.0%)	19.66 (4.81)	0.68 (0.11)	27.07 (21.23)
Moderately Overweight	141 (14.4%)	22.15 (5.74)	0.75 (0.14)	45.83 (26.40)
Slightly Overweight	251 (25.8%)	24.23 (5.09)	0.78 (0.13)	61.83 (26.33)
About the right Weight	418 (42.8%)	28.26 (5.16)	0.87 (0.13)	88.21 (19.83)
Slightly Underweight	59 (6.1%)	27.22 (5.40)	0.84 (0.14)	86.34 (16.37)
Moderately Underweight	7 (0.7%)	25.86 (4.67)	0.74 (0.15)	85.85 (15.68)
Very Underweight	2 (0.2%)	22.50 (6.36)	0.69 (0.44)	52.38 (6.06)
P - Value	-	<0.001	<0.001	<0.001
Self- Assessed Health				
Excellent	168 (17.2%)	29.95 (4.72)	0.92 (0.10)	89.28 (22.66)
Very Good	275 (28.2%)	27.58 (5.03)	0.86 (0.11)	80.43 (24.55)
Good	313 (32.1%)	23.72 (5.29)	0.77 (0.13)	59.54 (30.94)
Fair	157 (16.1%)	21.96 (5.36)	0.72 (0.14)	55.13 (29.59)
Poor	62 (6.4%)	20.48 (5.81)	0.68 (0.16)	46.62 (27.58)
P - Value	-	<0.001	<0.001	<0.001
Illness or Disability				
Yes	560 (57.4%)	23.06 (5.68)	0.75 (0.14)	58.46 (31.20)
No	415 (42.6%)	28.54 (4.85)	0.89 (0.11)	83.27 (23.89)
P - Value	-	<0.001	<0.001	<0.001
* As per Cole et al [33], individuals in the 85th percentile of the weight distribution for their age and gender were classed as overweight, and those in the 95th percentile were classed as being obese.				

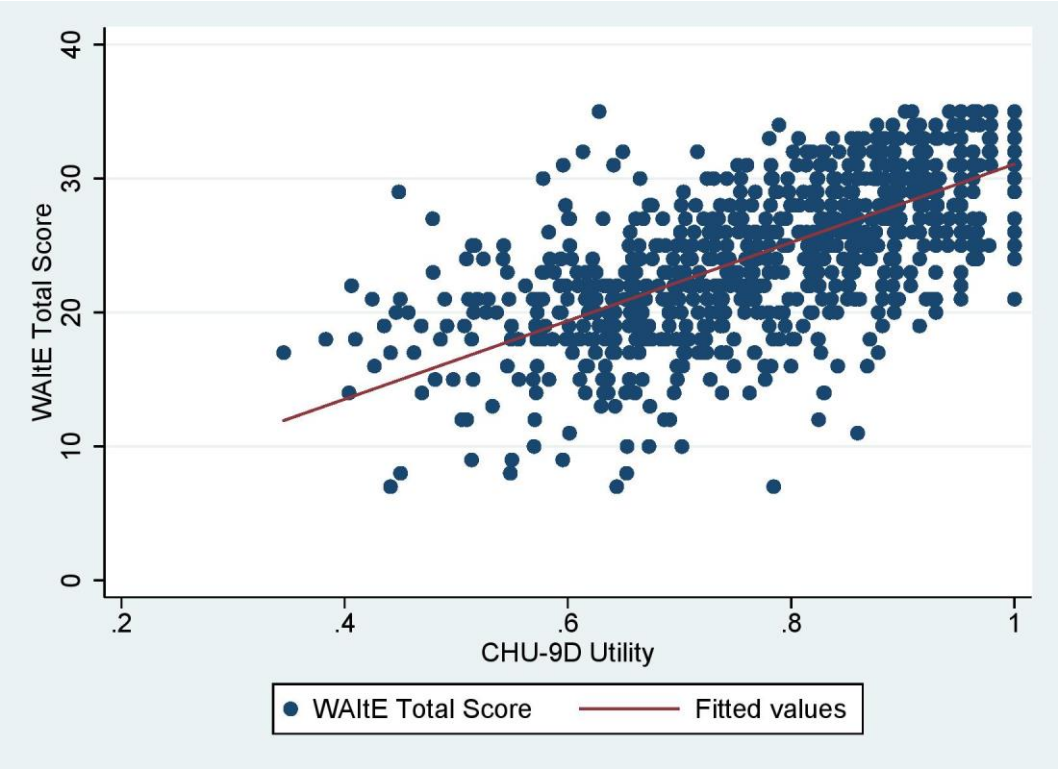


Figure 1- Correlation between the WAItE Total Score and the CHU-9D Utility Score

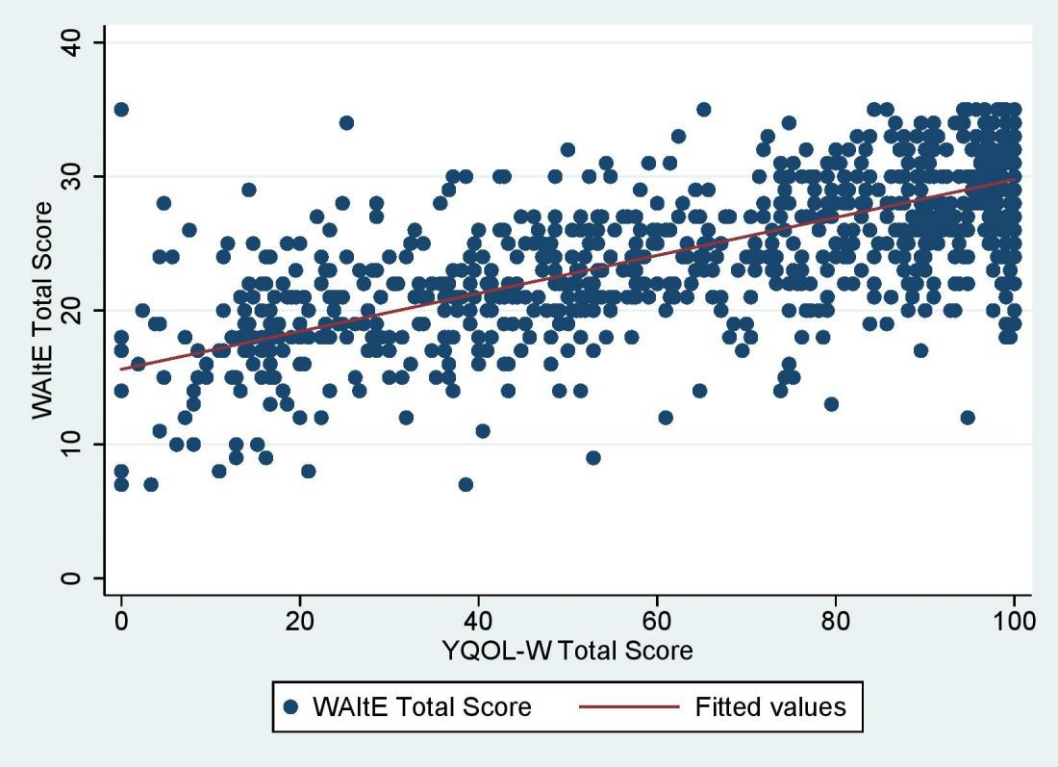


Figure 2- Correlation between the WAItE Total Score and the YQOL-W Total Score

Table 2- Intraclass Correlation Coefficients for the WAItE total score	
ICC Model	ICC (95% CI)
One way individual random-effects	0.795* (0.709, 0.858)
Two-way individual random-effects	0.796* (0.708, 0.859)
Two-way individual mixed-effects	0.801* (0.716, 0.862)
Notes: n = 97, * = p<0.01	

Table 3- Test-re-test correlations and percentage agreement for individual WAItE attributes				
Measure and item	**Test-Re-Test Correlation (Spearman's Rank)**	**Percentage Agreement (Weighted Kappa)**	**Weighted Kappa coefficient (Linear Weights)**	**Weighted Kappa coefficient (Quadratic Weights)**
Tired	0.584*	88.66*	0.517*	0.606*
Walking	0.678*	87.11*	0.521*	0.663*
Sports	0.811*	90.21*	0.710*	0.798*
Concentration	0.684*	86.86*	0.563*	0.665*
Embarrassed	0.661*	87.11*	0.590*	0.674*
Unhappy	0.691*	86.60*	0.488*	0.628*
Treated Different	0.608*	88.92*	0.522*	0.602*
Notes: n = 97, * = p<0.01				

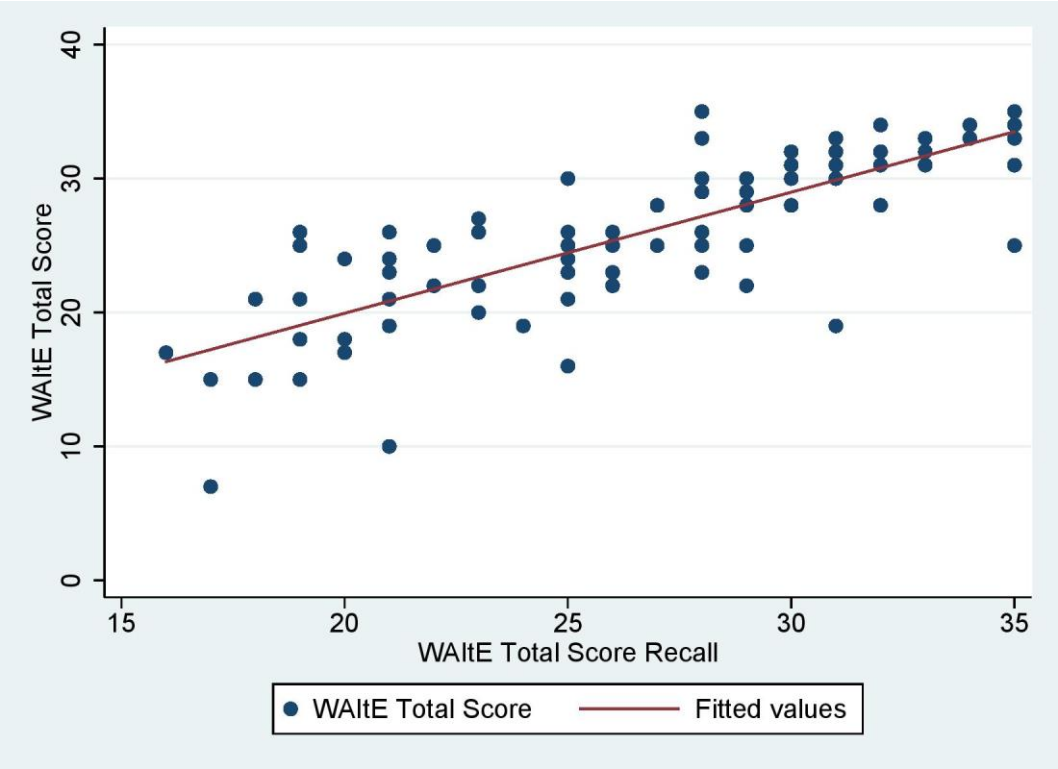
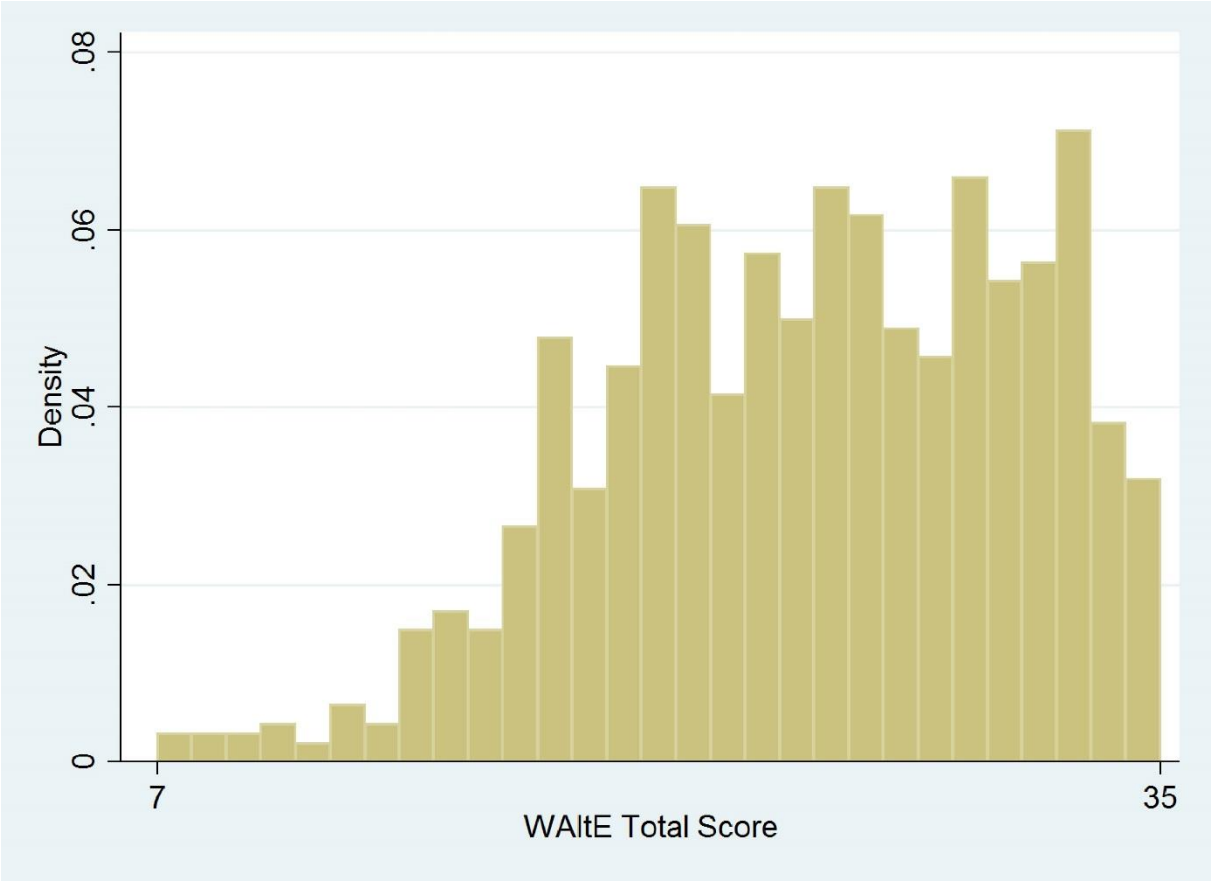


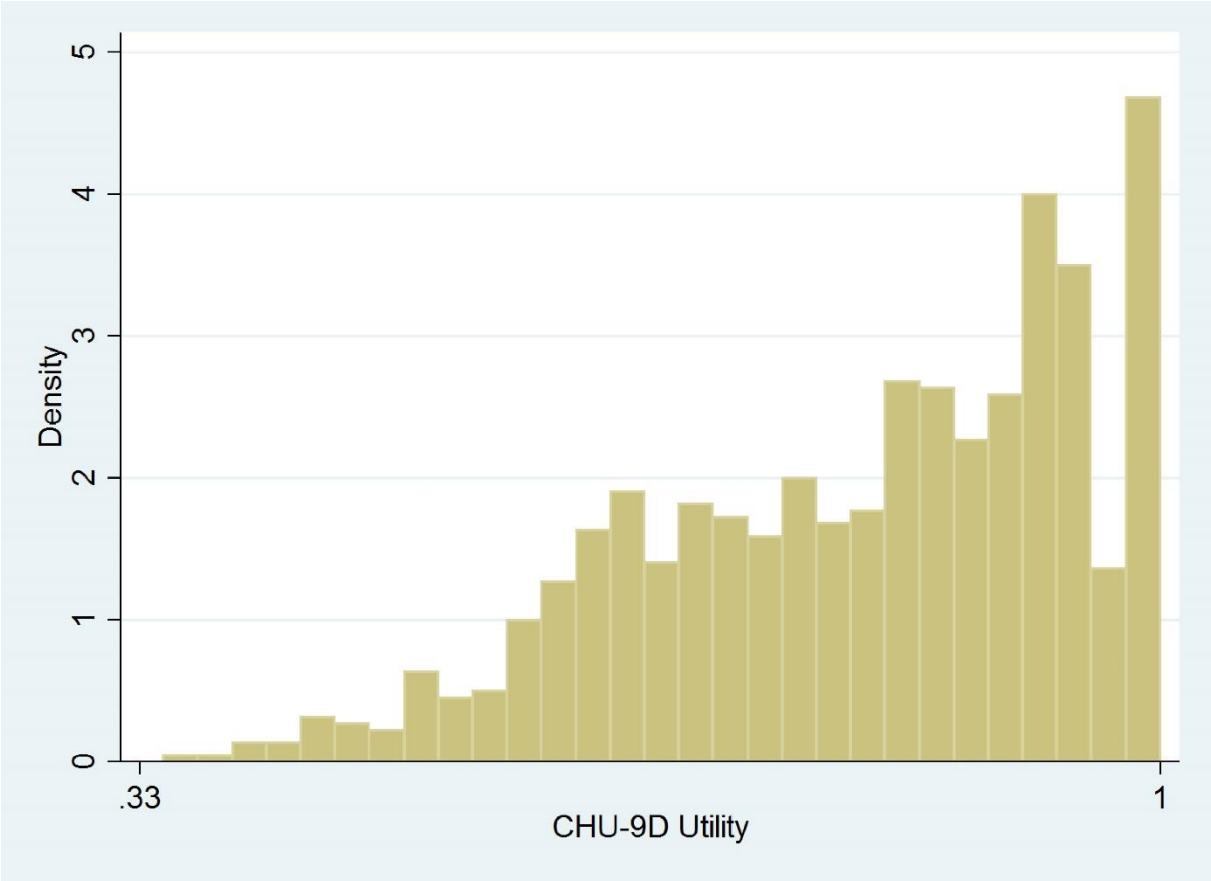
Figure 3- Correlation between the WAItE Total Score and the WAItE Total Score Recall

Online Supplementary Materials- Correlations between individual WAItE and CHU-9D items							
CHU-9D item	**WAItE item**						
	Tired	Walking	Sports	Concentrate	Embarrassed	Unhappy	Treated
Worry	0.319*	0.365*	0.314*	0.373*	0.403*	0.452*	0.397*
Sad	0.326*	0.277*	0.270*	0.352*	0.417*	0.416*	0.351*
Annoyed	0.251*	0.332*	0.268*	0.252*	0.287*	0.360*	0.338*
Tired	0.562*	0.371*	0.368*	0.406*	0.385*	0.413*	0.335*
Pain	0.291*	0.296*	0.231*	0.364*	0.316*	0.355*	0.300*
Sleep	0.226*	0.222*	0.228*	0.407*	0.317*	0.318*	0.268*
Daily	0.297*	0.330*	0.325*	0.407*	0.370*	0.362*	0.317*
Work	0.227*	0.376*	0.338*	0.413*	0.411*	0.463*	0.448*
Activity	0.347*	0.394*	0.437*	0.413*	0.373*	0.509*	0.445*
Notes: $n = 97$, * = $p<0.01$							

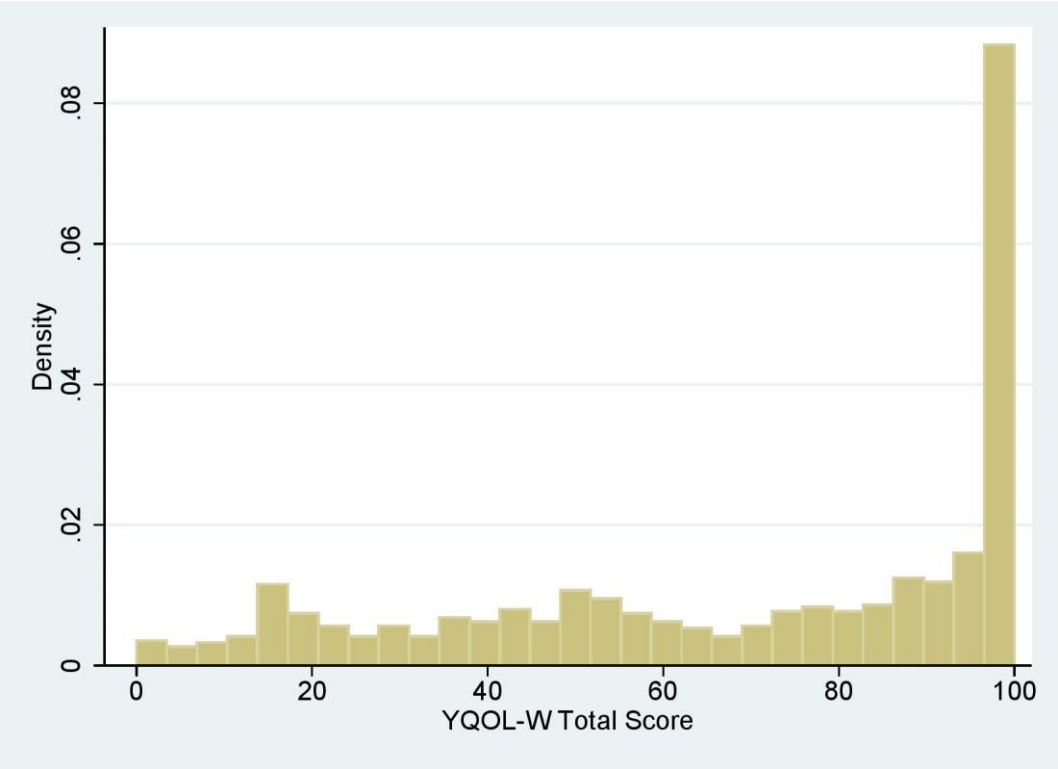
Online Supplementary Materials- Correlations between individual WAItE and YQOL-W items							
YQOL-W item	**WAItE item**						
	Tired	Walking	Sports	Concentrate	Embarrassed	Unhappy	Treated
Depressed	0.330	0.450	0.463	0.419	0.615	0.582	0.504
Ashamed	0.337	0.446	0.470	0.406	0.634	0.595	0.487
Uncomfortable	0.326	0.425	0.466	0.377	0.563	0.530	0.472
Clothes	0.344	0.432	0.497	0.417	0.633	0.599	0.498
Unattractive	0.333	0.449	0.466	0.414	0.611	0.571	0.533
Hide	0.284	0.451	0.464	0.376	0.650	0.568	0.494
Exercise	0.319	0.475	0.527	0.420	0.623	0.579	0.477
Eat	0.224	0.417	0.410	0.374	0.596	0.545	0.487
Notice	0.257	0.464	0.452	0.395	0.660	0.593	0.545
Worry	0.311	0.451	0.465	0.423	0.606	0.571	0.501
Social	0.284	0.460	0.464	0.402	0.654	0.603	0.533
Loser	0.298	0.450	0.436	0.389	0.578	0.572	0.512
Move	0.229	0.468	0.430	0.401	0.595	0.604	0.523
Swim	0.341	0.415	0.475	0.369	0.607	0.567	0.467
Girlfriend	0.225	0.357	0.370	0.383	0.525	0.526	0.481
Stare	0.228	0.451	0.397	0.366	0.614	0.588	0.561
Include	0.165	0.425	0.378	0.342	0.554	0.573	0.596
Job	0.161	0.430	0.374	0.346	0.563	0.563	0.568
Wear Clothes	0.307	0.456	0.440	0.403	0.646	0.621	0.506
Find Clothes	0.238	0.482	0.419	0.372	0.632	0.612	0.516
Exercising	0.261	0.502	0.492	0.387	0.579	0.613	0.538



Online Supplementary Materials- Distribution of the WAItE Total Score



Online Supplementary Materials- Distribution of the CHU-9D Utility Score



Online Supplementary Materials- Distribution of the YQOL-W Total Score